ORIGINALARBEIT

Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests

Kirsten Wagener, Thomas Brand, Birger Kollmeier
AG Medizinische Physik, Universität Oldenburg, D-26111 Oldenburg

Zusammenfassung *In Anlehnung an den schwedischen Satztest nach Hagerman (Hagerman 1984) wurde ein deutscher Satztest entwickelt (Wagener et al. 1999b). Jeder Satz dieses »Oldenburger Satztests« besteht aus fünf Wörtern, die jeweils zufällig aus 10 Alternativen ausgewählt wurden. In vorhergehenden Beiträgen (Wagener et al. 1999b, 1999a) wurde aufgrund von ersten Messungen mit der Auswahl von 10 Testlisten das Testinventar des Oldenburger Satztests festgelegt. In diesem Beitrag werden nun die theoretischen Erwartungen durch unabhängige Evaluationsmessungen mit Normalhörenden überprüft. Die Erwartungen werden größtenteils bestätigt: Der L_{50} (Signal-Rausch-Verhältnis, bei dem 50 % verstanden wurde) beträgt -7,1 dB S/N mit einer geringen Standardabweichung zwischen den Testlisten von 0,16 dB S/N. Der aufgrund von Messungen mit stärker trainierten Versuchspersonen erwartete Wert war -8,4 dB S/N (vgl. Wagener et al. 1999a). Die Steigung stimmt mit 17,1 %/dB (Standardabweichung 1,6 %/dB) mit der erwarteten von 17,2 %/dB überein. Die geringen Streuungen sowie der mit dem Friedman-Test nicht nachweisbare Unterschied zwischen den Listen bestätigen die perzeptive Äquivalenz der Testlisten. Die Vorhersagbarkeit der Sätze (ausgedrückt durch die Anzahl der statistisch unabhängigen Teile eines Satzes j) ist, wie aufgrund des von Hagerman angegebenen Wertes von $j \approx 4$ (Hagerman 1996) erwartet, sehr gering ($j = 4,29$ für -5 dB S/N und $j = 3,18$ für -9 dB S/N). Der Trainingseffekt beträgt 1 bis 2 dB S/N und kann durch Darbieten einer bis zwei Übungslisten auf weniger als 1 dB S/N begrenzt werden.*

Insgesamt bietet sich der Oldenburger Satztest daher als innovatives, valides und reliables Testverfahren für die klinische Audiologie an.

Schlüsselwörter: Sprachaudiometrie
Sprachverständlichkeit
Evaluation
Meßgenauigkeit

Corresponding author: Dipl.-Phys. Kirsten Wagener
AG Medizinische Physik, Carl von Ossietzky Universität Oldenburg
Carl-von-Ossietzky-Str. 9–11, D-26111 Oldenburg
Phone +49 441 798 3566, Fax +49 441 798 3698
E-mail: kirsten@medi.physik.uni-oldenburg.de

Development and evaluation of a German sentence test Part III: Evaluation of the Oldenburg sentence test

Kirsten Wagener, Thomas Brand, Birger Kollmeier
AG Medizinische Physik, Universität Oldenburg, D-26111 Oldenburg

Summary *A German sentence test described in companion papers (Wagener et al. 1999b, 1999a) is evaluated with respect to the performance in normal listeners, the redundancy of the material and the equivalence of the test lists. Similar to the Swedish test proposed by Hagerman (Hagerman 1984), each sentence of the »Oldenburger Satztest« is composed of a pseudo-random selection of five words taken from a list of 10 alternatives.*

This study compares theoretical expectations that are based on preliminary experiments and are described in Wagener et al. (1999a) with independent evaluation measurements with normal-hearing subjects. For the most part, our expectations were confirmed: the speech reception threshold L_{50} (speech level that corresponds to 50 % intelligibility) was -7.1 dB S/N with a small standard deviation of 0.16 dB across the test lists. The slope was to 17.1 %/dB (standard deviation: 1.6 %/dB) and matched the expected slope of 17.2 %/dB. The small standard deviations and lack of differences across lists (Friedman test) show the equivalence in intelligibility of the test lists. The redundancy of the sentences (described by number 3 of the statistically independent elements per sentence) was very low ($j = 4.29$ at -5 dB S/N and $j = 3.18$ at -9 dB S/N), which was expected because of the value $j \approx 4$ given by Hagerman (Hagerman 1996). The learning effect was 1–2 dB and can be reduced to less than 1 dB if one or two training lists are performed prior to data collection.

It may hence be concluded that the Oldenburg sentence test is an innovative, valid and reliable test procedure for audiology.

Keywords: *speech audiometry*
speech intelligibility
evaluation
precision of measurement

Einleitung

Der in den vorhergehenden Beiträgen (*Wagener* et al. 1999b, 1999a) vorgestellte Oldenburger Satztest soll eine Lücke in der Sprachaudiometrie schließen und ein effizientes, nicht von der Wahl der Testliste abhängiges Verfahren für die Bestimmung der Sprachverständlichkeit im Störgeräusch mit einer großen Anzahl an wiederholbaren Testlisten bereitstellen. In Anlehnung an den schwedischen Satztest nach *Hagerman* (*Hagerman* 1984) besteht jeder Satz der Form Name-Verb-Zahlwort-Adjektiv-Objekt aus einer (pseudo-) zufälligen Auswahl der einzelnen Testwörter, für die jeweils 10 Alternativen zur Verfügung stehen. Die Phonemverteilung des Sprachmaterials entspricht der mittleren Phonemverteilung der deutschen Sprache (vgl. *Wagener* et al. 1999b). Im Gegensatz zu dem Marburger Satztest (*Niemeyer* 1967) und dem Göttinger Satztest (*Kollmeier* und *Wesselkamp* 1997) kann der Test daher beliebig oft mit derselben Versuchsperson wiederholt durchgeführt werden. In diesem Artikel werden die wesentlichen Eigenschaften des Tests mit einem unabhängigen Versuchspersonenkollektiv untersucht und mit den aus theoretischen Überlegungen und vorherigen Messungen getroffenen Erwartungen verglichen (Vergleichbarkeit der Testlisten, Form der Diskriminationsfunktion, Vorhersagbarkeit der Sätze). Dadurch soll geklärt werden, ob der Test den an ihn gestellten Anforderungen (siehe oben sowie *Wagener* et al. 1999b) genügt.

Die Messungen zur Optimierung der Testlisten (*Wagener* et al. 1999a) sind mit hochgradig trainierten Versuchspersonen durchgeführt worden, die Homogenität der Listen soll jedoch auch bei den klinischen Anwendungen mit naiven Probanden gewährleistet sein. Dies ist von großer Bedeutung für die praktische Anwendbarkeit des Tests, weil die Testergebnisse weitestgehend unabhängig von der jeweils eingesetzten Testliste immer denselben Wert liefern sollen. Aufgrund der vorherigen Messungen wurden die Listen so zusammengestellt, daß diese Voraussetzung erfüllt ist. Um dieselbe Aussage auch für ein unabhängiges, der klinischen Population eher entsprechendes Versuchspersonenkollektiv machen zu können, werden die Testlisten mit einer Gruppe größtenteils in Sprachverständlichkeitsmessungen unerfahrenen Versuchspersonen evaluiert. Dadurch können Normwerte für die praktische audiologische Anwendung gegeben werden.

In der Literatur sind Evaluationsmessungen mit naiven Probanden nur für den Göttinger Satztest (*Kollmeier* und *Wesselkamp* 1997) beschrieben. Bei anderen Tests, wie z. B. dem Freiburger oder Marburger Test, wurde die perzeptive Äquivalenz der Listen nicht evaluiert. Das hat zur Folge, daß die gemessene Verständlichkeit von der verwendeten Testliste abhängt und so das Meßergebnis durch die Wahl der Testliste beeinflußt werden kann.

Im folgenden werden die Evaluationsmessungen mit 20 Normalhörenden vorgestellt. Der Trainingseffekt wird zunächst durch adaptive Messungen des L_{50} (d. h. des zu 50 % Sprachverständlichkeit gehörenden Signal-Rausch-Verhältnisses) bestimmt. Durch darauffolgende Messungen bei konstanten Signal-Rausch-Verhältnissen werden die Verständlichkeiten der einzelnen Testlisten ermittelt, um die Äquivalenz der Listen zu überprüfen. Zusätzlich werden die Messungen in Bezug auf die Vorhersagbarkeit des Satzmaterials ausgewertet. Die Evaluationsmessungen wurden mit den 12 in *Wagener* et al. (1999a) ausgewählten Listen durchgeführt (10 Testlisten und 2 Trainingslisten). Im Folgenden werden die Auswertungen jedoch nur für die 10 Testlisten durchgeführt. Lediglich bei der Messung des Trainingseffekts werden – bedingt durch das adaptive Meßverfahren – alle 12 gemessenen Listen berücksichtigt.

Methode

Die 12 Testlisten à 10 Sätze wurden mit 20 normalhörenden Versuchspersonen evaluiert. Die Listen wurden zur Verkürzung der Meßzeit in sechs Doppellisten aus 20 Sätzen zusammengefaßt, dieses Vorgehen führt zu einer Meßzeit von ca. 5 min pro Doppelliste. Die Messungen wurden in einer schallisolierten Hörkabine mit einer im Rahmen eines Verbundprojekts zur Sprachaudiometrie entwickelten Apparatur durchgeführt (*Kollmeier* et al. 1992): Über einen Pentium PC mit einer Ariel DSP 32C-Karte (mit 16 bit AD-DA-Wandlern) wurde der gesamte Meßvorgang gesteuert. Die Sprachsignale sowie das Störgeräusch lagen digital auf einer Festplatte vor (mit 25 kHz Samplingfrequenz und 16 bit Auflösung). Die Darbietungspegel wurden über ein computergesteuertes Audiometer (im Rahmen des oben erwähnten Projekts entwickelt, siehe *Kollmeier* et al. 1992) eingestellt, über den Signalprozessor wurde das Sprachsignal und das Rauschen im gewünschten Signal-Rausch-Verhältnis gemischt. Die Testsignale wurden den Versuchspersonen diotisch über einen breitbandig kalibrierten Kopfhörer des Typs Sennheiser HDA 200 dargeboten. Die Probanden hatten die Aufgabe, die von ihnen verstandenen Sätze oder Satzteile dem Versuchsleiter zu wiederholen. Dieser markierte jedes falsch wiedergegebene Wort auf dem berührungsempfindlichen Bildschirm eines Handheld-Computers Epson ETH10S (dort wurden alle Wörter des Satzes dargestellt). Über eine serielle Schnittstelle wurden die Antworten an den Meßcomputer weitergeleitet und dort für die weiteren Auswertungen gespeichert. Als Störgeräusch wurde das von *Wagener* et al. (1999b) beschriebene »Oldenburger Rauschen« bei einem konstanten Pegel von 65 dB SPL verwendet.

An den Messungen nahmen 20 Versuchspersonen (14 Männer, 6 Frauen) im Alter von 23 bis 42 Jahren (mittleres Alter 29 Jahre) teil. Sie zeigten aufgrund des Tonaudiogramms und ihrer Höranamnese keine klinischen Auffälligkeiten. Vier dieser Pro-

banden waren ähnlichen Testsituationen vertraut (Mitglieder der Arbeitsgruppe »Medizinische Physik« an der Carl von Ossietzky-Universität Oldenburg), die restlichen (bezahlten) Personen waren auf diesem Gebiet unerfahren.

Um bei allen Versuchspersonen den gleichen Trainingsgrad zu erzielen und gleichzeitig Informationen über den Trainingseffekt zu erhalten, wurde zunächst innerhalb einer halben Stunde der L_{50}-Wert jeder Doppelliste adaptiv gemessen. Es wurde dabei als adaptive Pegelsteuerung das von *Brand* (1994) verallgemeinerte Verfahren nach *Hagerman* und *Kinnefors* (1993) verwendet. Die Schrittweite zum nächstfolgenden Darbietungspegel wird aus der vorangehenden Antwort wie folgt berechnet: $\Delta L = -\frac{f_i(SV-T)}{m}$, wobei *SV* die Satzverständlichkeit des vorhergehenden Satzes, *T* die Ziel-Richtig-Antwort-Wahrscheinlichkeit (in diesem Fall 50 %), *m* die Steigung und f_i die »Geschwindigkeit« der Steuerung ist, diese beträgt zu Beginn der Messung 2, halbiert sich nach dem 1. Wendepunkt zu 1 und nach dem 2. Wendepunkt zu 0,5.

Die Listen wurden den Versuchspersonen in unterschiedlicher Reihenfolge dargeboten. Nach einer ca. 10 bis 15 min Pause wurden die eigentlichen Evaluationsmessungen durchgeführt.

Die 20 Personen wurden in zwei Gruppen unterteilt. Der einen Gruppe wurden die Testlisten 1 bis 6 (Doppellisten 1 bis 3) bei einem konstanten Signal-Rausch-Verhältnis von -5 dB S/N und die Listen 7 bis 12 (Doppellisten 4 bis 6) bei -9 dB S/N dargeboten, der anderen Gruppe genau umgekehrt. Die Listen wurden jeweils in unterschiedlicher Reihenfolge gemessen, die beiden Signal-Rausch-Abstände wurden immer abwechselnd dargeboten. Weiter ausgewertet wurden nur die 10 Testlisten, obwohl auch die 2 Trainingslisten mit gemessen wurden (diese wurden jedoch aus dem Testinventar wegen zu hoher Standardabweichung der wortspezifischen L_{50}-Werte gestrichen). Es wurde jeweils die Verständlichkeit der gesamten Liste sowie der einzelnen Sätze und Wörter bestimmt, um Aussagen über die Eigenschaften der Testlisten zu erhalten und die Vorhersagbarkeit der Sätze (als Maß wird der *j* Faktor verwendet: $j = \frac{log(p_s)}{log(p_w)}$; p_s :Wahrscheinlichkeit, daß ein Satz komplett verstanden wurde, p_w: Wahrscheinlichkeit, daß ein Wort richtig verstanden wurde; siehe auch *Boothroyd* und *Nittrouer* 1988) zu bestimmen.

Ergebnisse

Trainingseffekt

Die adaptiven L_{50}-Messungen, die vor den Evaluationsmessungen durchgeführt wurden, um einen einheitlichen Trainingsgrad aller Versuchspersonen zu erhalten, wurden zur Abschätzung des Trainingseffekts ausgewertet. In Abbildung 1 ist der über alle 12 Versuchspersonen arithmetisch gemittelte L_{50}-Wert mit der Standardabweichung über der Nummer der Messung aufgetragen. Die Nummer identifiziert nicht die Testliste, sondern die Reihenfolge der Messung (Messung Nr. 1 beinhaltet alle Listen, die am Anfang gemessen wurden). Zusätzlich ist der mittlere L_{50} der darauffolgenden Evaluationsmessungen angegeben. Da die Listen immer in unterschiedlicher Reihenfolge dargeboten wurden, sind die L_{50}-Werte aus Abb. 1 unabhängig von möglichen Unterschieden zwischen den Testlisten. Das zeigt auch Abbildung 2, auf der die mittleren L_{50}-Werte der adaptiven Messungen über den einzelnen Testlisten dargestellt sind. Da der Trainingseffekt in diese Messungen eingeht, resultieren relativ große Standardabweichungen. Zwischen den ersten beiden dargebotenen Listen ergibt sich aus Abb. 1 ein Unterschied im L_{50} von 1 dB S/N, die Differenzen zwischen den darauffolgenden Listen sind jeweils kleiner als 0,5 dB S/N. Dies entspricht etwa der Genauigkeit, mit der die Verständlichkeitsschwelle im adaptiven Verfahren bestimmt wird. Dieser Trainingseffekt kann durch Gewöhnung der Versuchspersonen an das Meßverfahren erklärt werden, da sie größtenteils vor diesem Experiment an keinen Sprachverständlichkeitsmessungen teilgenommen haben. Nach zwei gemessenen Listen ist ihnen der formale Aufbau der Sätze bewußt (fünf Wörter, Satzbau: Name-Verb-Zahl-Adjektiv-Objekt), was ebenfalls zu einer besseren Verständlichkeit führt. Insgesamt zeigt sich über alle sechs Messungen hinweg ein Lerneffekt von 1 bis 2 dB S/N, der im wesentlichen während der ersten beiden Listen stattfindet. Die Abweichungen des Unterschieds zwischen 5. und 6. Messung vom Unterschied zwischen der 3. und 4. sowie 4. und 5. Messung sind jedoch nicht signifikant. Der mittlere L_{50}der Testlisten, der durch die Evaluationsmessungen bestimmt wurde (siehe unten), beträgt -7,1 dB S/N. Damit ist er gleich dem mittleren L_{50} der 6. adaptiven Messung. Über die 6 Messungen mit konstantem Signal-Rausch-Verhältnis ergab sich demnach kein weiterer Lerneffekt. Damit kann der Lerneffekt mit maximal 2 dB S/N nach oben hin abgeschätzt werden, 1 dB S/N davon treten innerhalb der ersten beiden Testdurchgänge auf. Um einen Trainingseffekt sicher auszuschließen, sollten in der Praxis jeweils ein bis zwei Übungslisten (d. h. je nach Genauigkeitsanforderungen bis zu 60 Sätze) vorweg gemessen werden.

Die Messungen zur Optimierung des Testmaterials aus *Wagener* et al. (1999a) ergaben einen mittleren L_{50} von -8,4 dB S/N. Er liegt um ca. 1 dB S/N niedriger als der L_{50} der bei den adaptiven Messungen zuletzt gemessene. Der Unterschied kann darauf zurückgeführt werden, daß die Versuchspersonen zur Bestimmung der wortspezifischen Diskriminationsfunktionen (Mitglieder der Arbeitsgruppe) in Sprachverständlichkeitsmessungen hochgradig trainiert waren und das verwendete Wortmaterial kannten.

Unterschiede in der Verständlichkeit der Testlisten

Um die in *Wagener* et al. (1999a) theoretisch berechneten Eigenschaften der Testlisten (L_{50} und m_{ges}, Signal-Rausch-Verhält-

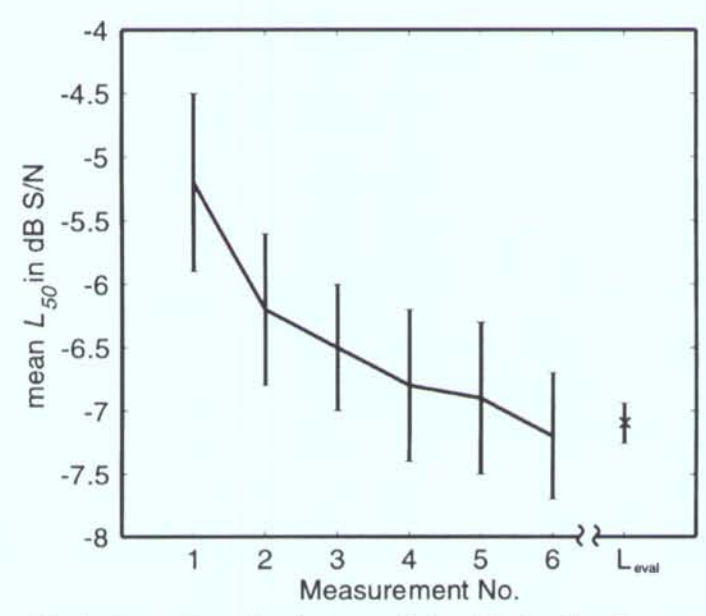


Abb. 1: Darstellung des Trainingseffekts: Mittlere L_{50}-Werte aller Versuchspersonen (mit Standardabweichungen) in Abhängigkeit von der Darbietungsreihenfolge, L_{eval} zeigt den mittleren L_{50}-Wert der Evaluationsmessungen.

Fig. 1: Assessment of the training effect: Mean speech reception thresholds of all subjects (L_{50}, with standard deviations) as a function of the temporal order of performing the measurements. The abcissa denotes the number of the adaptive measurement track (using a double list of 20 sentences each) performed as practice runs prior to the evaluation measurements. L_{eval} denotes the mean L_{50} of the subsequent evaluation measurements. Note that each subject was trained with a different sequence of test lists.

nis, bei dem 50 % verstanden wurde und Steigung der Gesamt-Diskriminationsfunktion am L_{50}) praktisch zu überprüfen, wurden die Evaluationsmessungen wie folgt ausgewertet:

Für jede Testliste wurden die Verständlichkeiten aller 10 Versuchspersonen pro Signal-Rausch-Verhältnis gemittelt. Daraus resultierten für jede Liste zwei Meßpunkte bei einer Darbietung von -5 und -9 dB S/N. Durch Einsetzen dieser beiden Punkte in die logistische Funktion, die als Modellfunktion die Diskriminationsfunktion (Abhängigkeit der Verständlichkeit vom Signal-Rausch-Verhältnis) nachbildet: $f(x) = \frac{1}{1+exp(-\frac{x-L_{50}}{s})}$ (vgl. *Wagener* et al. 1999a) erhält man die zugehörenden Parameter L_{50} und $m_{ges} = \frac{1}{4s}$ der Testliste. Tabelle 1 zeigt die so erhaltenen Ergebnisse.

Für die 10 Testlisten wird erwartet, daß sie hinsichtlich ihrer Sprachverständlichkeit äquivalent sind. Um diese Äquivalenz nachzuweisen, wird ein indirektes Verfahren verwendet: Die Meßergebnisse wurden für die beiden Signal-Rauschabstände -5 und -9 dB S/N einer Rangvarianzanalyse nach Friedman (*Sachs* 1992) unterzogen. Zum Prüfen der Nullhypothese (alle Bedingungen entstammen einer Grundgesamtheit: Alle Testlisten liefern äquivalente Ergebnisse bei der Messung der Sprachverständlichkeit) hat Friedman die Prüfgröße $\hat{\chi}^2_R$ angegeben. Bei Erreichen oder Überschreiten der für gegebene n Stichproben und k Bedingungen tabellierten Schrankenwerte χ^2_R kann die Nullhypothese mit einer bestimmten Irrtumswahrscheinlichkeit abgelehnt werden.

Die 20 Probanden waren in zwei Gruppen aufgeteilt, sechs Testlisten wurden von der einen Gruppe bei einem Signal-Rausch-Verhältnis von -5 dB S/N und von der anderen bei -9 dB S/N gemessen, vier Listen bei -9 dB S/N und -5 dB S/N. Daher muss der Friedman-Test viermal durchgeführt werden. Der Schrankenwert für $n = 10$ Stichproben und $k = 6$ Bedingungen bei 5 % Irrtumswahrscheinlichkeit ist χ^2_R = 10,76. Die errechneten Prüfgrößen betragen für die ersten sechs Testlisten bei einem Signal-Rausch-Verhältnis von -5 dB S/N $\hat{\chi}^2_R$ =4,47 und für -9 dB S/N $\hat{\chi}^2_R$ =4,01. Bei $n = 10$ und $k = 4$ ist χ^2_R =7,67, errechnet wurden dagegen $\hat{\chi}^2_R$ = 3 für -5 dB S/N und $\hat{\chi}^2_R$ = 3,96 bei einem Signal-Rausch-Verhältnis von -9 dB S/N.

Damit liegen die Prüfgrößen jeweils sehr deutlich unter den angegebenen Schrankenwerten für eine Irrtumswahrscheinlichkeit von 5 %. Die Äquivalenz der Testlisten kann daher auf dem 5 %-Niveau nicht abgelehnt werden.

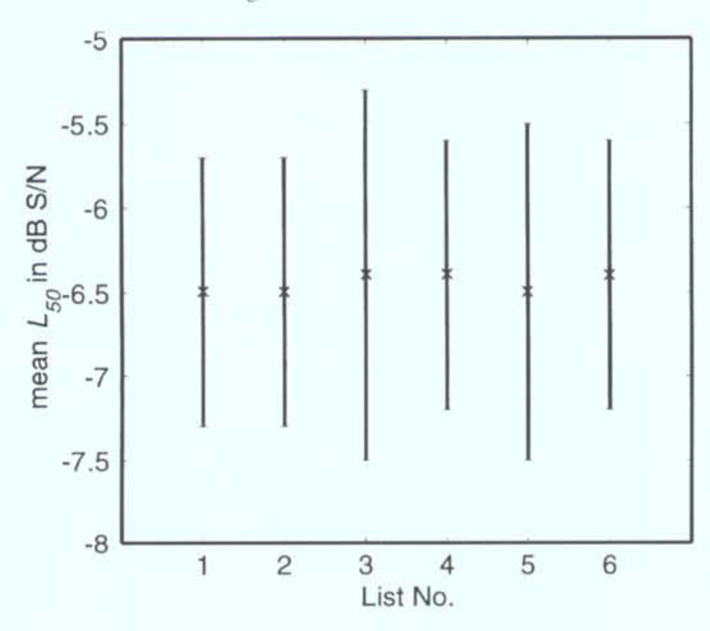


Abb. 2: L_{50}-Werte über alle Versuchspersonen gemittelt, sowie zugehörige Standardabweichungen aus den adaptiven Messungen zur Bestimmung des Trainingeffektes in Abhängigkeit von den einzelnen Doppellisten.

Fig. 2: Mean L_{50}-values and standard deviations of all subjects as a function of the double test list number. The same data as in Fig. 1 are employed (measured by an adaptive procedure), but sorted in a different way.

Testliste	SV [%] bei -9 dB S/N	SV [%] bei -5 dB S/N	Berechnet L_{50} [dB S/N]	Berechnet m [%/dB]
1	24,2	78,8	-7,1	15,3
2	21,0	81,2	-7,1	17,4
3	23,4	79,4	-7,1	15,8
4	23,8	78,8	-7,1	15,5
5	21,6	78,2	7,0	16,0
6	27,2	84,4	-7,5	16,7
7	19,0	80,2	-7,0	17,8
8	18,2	82,8	-7,0	19,2
9	16,6	83,6	-7,0	20,3
10	21,6	79,4	-7,0	16,5
Mittelwert	21,7	80,7	-7,1	17,1
Standardabweichung	3,20	2,21	0,16	1,65

Tab. 1: Verständlichkeiten (SV) der beschriebenen Evaluationsmessungen sowie daraus errechnete Parameter der Diskriminationsfunktionen für die einzelnen Testlisten.

Table 1: Speech intelligibility obtained from each test list and 20 normal hearing subjects at the two different signal-to-noise ratios -9 dB and -5 dB. Based on these values, the parameters L_{50} (i. e. signal-to-noise ratio corresponding to 50 % intelligibility) and m (i. e. slope at L_{50}) were computed and also listed. The last row gives the average and standard deviation across lists.

Mittelwert von L_{50} und m

Für die 10 einzelnen Testlisten wurden die Gesamt-Diskriminationsfunktionen unter Berücksichtigung der L_{50}-Verteilung der Einzelwörter nach dem Pegelangleich berechnet. Abbildung 3 zeigt den Vergleich dieser Funktionen mit den über die Versuchspersonen gemittelten Meßwerten. Wie in Abbildung 3 deutlich zu sehen, ist der mittlere L_{50}-Wert der Evaluationsmessungen mit -7,1 dB S/N um 1,3 dB S/N höher als der bei der Optimierung des Testmaterials (siehe *Wagener* et al. 1999a) gemessene und für die theoretischen Berechnungen verwendete (-8,4 dB S/N). Dies ist auf das geringere Training der Versuchspersonen im Vergleich zu den hochgradig mit derartigen Messungen trainierten Versuchspersonen aus den Optimierungsmessungen zurückzuführen.

Wird dieser Unterschied durch lineares Verschieben der Meßwerte um -1,3 dB S/N auf der x-Achse ausgeglichen, so erhält man Bild 4. Dieser Ausgleich bedeutet die Erhöhung des Signal-Rausch-Verhältnisses des gesamten Tests um 1,3 dB S/N, die relativen Zusammenhänge wie z. B. die Steigung der Diskriminationsfunktion werden durch dieses Vorgehen nicht verändert. Die aufgrund der Optimierungsmessungen theoretisch erwartete Steigung der Diskriminationsfunktionen von m = 17,2 %/dB (vgl. *Wagener* et al. 1999a) stimmt hervorragend mit der mittleren gemessenen Steigung von m = 17,1 %/dB überein (siehe auch Abb. 4). Die Erwartungen bezüglich anderer Streuparameter (z. B. Verständlichkeits- und Steigungsstreuung der Einzelwörter) konnte wegen zu weniger Meßpunkte durch die Evaluationsmessungen allerdings nicht überprüft werden.

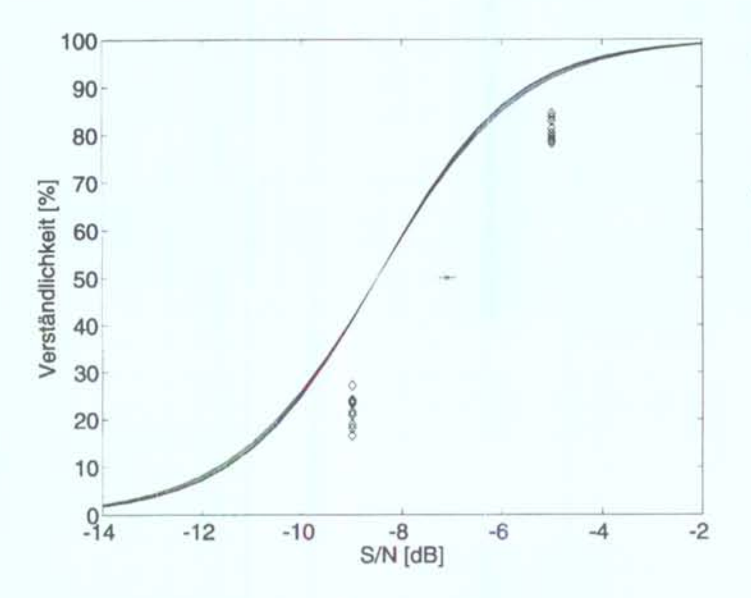


Abb. 3: Aufgrund der Optimierungsmessungen theoretisch berechnete Diskriminationsfunktionen aller 10 Testlisten (vgl. Wagener et al. 1999a) im Vergleich zu den Evaluationsmessungen: Die Diamanten kennzeichnen die über die 20 Versuchspersonen gemittelten Meßwerte für Jede der 10 Testlisten. Das Kreuz bezeichnet den daraus ermittelten mittleren L_{50}-Wert der Evaluationsmessungen mit zugehörender Standardabweichung.

Fig. 3: Expected discrimination functions for all 10 test lists based on the optimization measurements described by Wagener et al. (1999a). In comparison, the results of the current evaluation measurements are given: the diamonds denote the mean values of 20 normal-hearing subjects for each of the 10 lists. The cross denotes the mean L_{50} value of the evaluation measurements with its standard deviation.

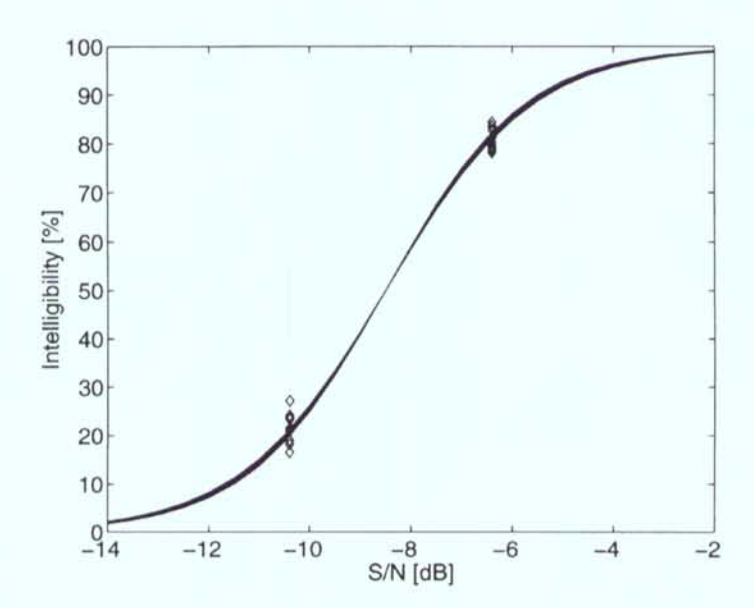


Abb. 4: Wie Abb. 3 nach dem Ausgleich der unterschiedlichen Trainingssituation durch Verschieben der »erwarteten« Diskriminationsfunktion.

Fig. 4: Same as figure 3 taking a training effect of 1.3 dB into account and hence shifting the »expected« discrimination function by this amount.

Anzahl der statistisch unabhängigen Satzteile, *j* Faktor

Boothroyd und *Nittrouer* führten als ein Maß für die Vorhersagbarkeit von Sätzen den *j* Faktor ein (*Boothroyd* und *Nittrouer* 1988), der die Anzahl statistisch unabhängiger Teile (in diesem Fall statistisch unabhängige Wörter pro Satz) beschreibt: Über den Zusammenhang $j = \frac{\log(p_s)}{\log(p_w)}$ wird *j* aus der Wahrscheinlichkeit p_s, daß ein Satz komplett verstanden wurde und p_w, daß ein Wort richtig verstanden wurde, berechnet.

Ein großer *j* Faktor ist insbesondere für die adaptive Teststeuerung wichtig, da bei einem Trial ein größerer Informationsgewinn stattfindet als bei kleinem *j* Faktor. Bei vorgegebener Zeit werden mehr Informationen erhalten und somit ist die Meßgenauigkeit höher. Bei den meisten anderen Satztests, wie z. B. dem Göttinger Satztest, liegt der *j* Faktor bei etwa 2.

Aus den durch die Evaluation gewonnenen Meßdaten wurde der *j* Faktor des Wortmaterials bei den Signal-Rausch-Verhältnissen von -5 und -9 dB S/N berechnet. Es ergab sich $j = 4{,}29$ bei einem Signal-Rausch-Verhältnis von -5 dB S/N (Verständlichkeit: 80,7 %) und $j = 3{,}18$ für -9 dB S/N (Verständlichkeit: 21,7 %). Die Vorhersagbarkeit der Sätze des Oldenburger Satztests entspricht denen des Satztests nach *Hagerman*: Für den schwedischen Test beträgt $j = 2{,}92$ bei einer Verständlichkeit von weniger als 28 % und $j = 4{,}77$ bei mehr als 82 % Verständlichkeit (*Hagerman* 1996). Die Sätze sind somit nicht so vorhersagbar wie die des Göttinger Satztests ($j = 2{,}38$ bei -4 dB S/N und $j = 1{,}95$ bei -8 dB S/N, vgl. *Kollmeier* und *Wesselkamp* 1997). Die Werte der *j* Faktoren werden an gleichen Stellen auf den Diskriminationsfunktionen verglichen, daher muß der Abstand zum jeweiligen L_{50}-Wert (beim Oldenburger Satztest: $L_{50} = -7{,}1$ dB S/N, beim Göttinger Satztest: $L_{50} = -6{,}1$ dB S/N) gleich sein.

Diskussion

Die Evaluation des Oldenburger Satztests mit einem unabhängigen Versuchspersonenkollektiv von Normalhörenden zeigte einen Trainingseffekt von maximal 2 dB S/N für untrainierte Versuchspersonen während einer etwa halbstündigen Trainingssitzung mit 12 adaptiv gesteuerten Testlisten.

Die mittlere Sprachverständlichkeitsschwelle L_{50} der sich anschließenden Evaluationsmessungen kann als stationär angenommen werden, weil sie nicht niedriger liegt als der nach 12 Trainingslisten (bzw. 6 Doppellisten) von den Versuchspersonen erreichte mittlere Pegel und weil sich während der Evaluationsmessungen der L_{50}-Wert nicht mehr signifikant verschoben hat. Dieser mittlere L_{50} liegt für das hier verwendete Versuchspersonenkollektiv 1,3 dB S/N über dem Wert für hochgradig trainierte Probanden aus den von *Wagener* et al. (1999a) vorgestellten Optimierungsmessungen.

Die Steigung der einzelnen Testlisten von im Mittel 17 %/dB entspricht exakt den Erwartungen für die Steigung der Gesamt-Diskriminationsfunktion, die mit dem probabilitischen Modell (*Kollmeier* et al. 1992) aus der Einzelwort-Diskriminationsfunktion und der Verteilungsfunktion der wortspezifischen L_{50}-Werte errechnet wurde (*Wagener* et al. 1999a).

Die Homogenität der Testlisten entspricht ebenfalls den Erwartungen, dies wurde durch die geringen Standardabweichungen zwischen den Testlisten von L_{50} (-7,1 dB S/N±0,16 dB S/N) und Steigung (17,1 %/dB±1,6 %/dB) sowie durch nicht signifikante Unterschiede im Friedman-Test gezeigt.

Der Trainingseffekt wurde mit einem größtenteils ungeübten und mit akustischen Fragestellungen nicht vertrauten Versuchspersonenkollektiv abgeschätzt. Dieses Kollektiv ist für die klinische Population repräsentativer als das Kollektiv der vorherigen Messungen, jedoch ist das Durchschnittsalter sicher niedriger als in der Klinik.

Bei der hier eingehaltenen Meßgenauigkeit von 0,5 dB S/N für den L_{50} liegt der Gewöhnungs- und Trainingseffekt im Bereich des 2 bis 4fachen der Genauigkeit, so daß eine genaue Schätzung des maximalen Trainingseffekts durchaus möglich ist.

Der Unterschied der stationären Schwelle von naiven und hochgradig trainierten Versuchspersonen läßt sich zum Teil dadurch erklären, daß die trainierten Probanden das Wortmaterial gut kannten. Die Messungen ähnelten daher einem geschlossenen Testverfahren, d. h. die Antworten können aus bestimmten Alternativen gewählt werden. Das bedeutet eine Beeinflussung des Ergebnisses durch die Ratewahrscheinlichkeit, die hier 10 % beträgt (es gibt 10 Alternativen für jedes Wort).

Für trainierte Versuchspersonen wird durch die Annahme eines geschlossenen Testverfahrens der Bereich der Verständlichkeit (beim offenen Test 0 bis 100 %) auf 10 bis 100 % gestaucht. Eine Verständlichkeit von 50 % beim geschlossenen Verfahren entspricht daher im offenen Verfahren einer Verständlichkeit von 45 %. Soll der unterschiedliche L_{50}-Wert auf diese Art erklärt werden, so müßte beim Einsetzen des Signal-Rausch-Verhältnisses von L = -8,4 dB S/N in die Funktion $f(L) \frac{1}{1+exp(-4 \cdot 0{,}171 \cdot (L+7{,}1))}$ (Modellfunktion aus *Wagener* et al. 1999a) mit den gemittelten Parametern der Testlisten: L_{50}=-7,1 dB S/N, $\frac{1}{s}=4 \cdot m_{ges}=4 \cdot 0{,}171 \frac{1}{dB}$) eine Verständlichkeit von 45 % resultieren. Es ergibt sich jedoch nur eine Verständlichkeit von ca. 32 %. Allein durch die Annahme eines geschlossenen Testverfahrens für die geübten Versuchspersonen kann der Unterschied im L_{50} demnach nicht erklärt werden.

Eine weitere Erklärung der besseren Schwellenwerte für die trainierte Probandengruppe kann die besondere »Hörerfahrung« dieser Personen geben. Die meisten waren Mitglieder der Arbeitsgruppe »Medizinische Physik« der Carl von Ossietzky-Universität Oldenburg, die mit vielen Formen von akustischen Experimenten vertraut sind. Die anderen trainierten Teilnehmer sind aufgrund von musikalischen Erfahrungen geprägt. Diese »Schärfung« des Gehörs scheint eine recht große Rolle zu spielen, denn ein Teilnehmer der Evaluationsmessungen, der Erfahrung mit akustischen Experimenten und im besonderen auch mit Sprachwahrnehmung hat, zeigte trotz gleichen Trainings signifikant bessere Schwellenwerte als die naiven Teilnehmer.

Nimmt man für die hochgradig trainierte Versuchspersonengruppe allgemein konstant bessere Schwellenwerte an als für die ungeübten Teilnehmer der Evaluationsmessungen, so kann die Übereinstimmung der Meßwerte mit den Erwartungen festgestellt werden, indem die Diskriminationsfunktion um diesen Schwellenunterschied verschoben wird.

Eine dritte Einflußgröße auf den beobachteten Unterschied zwischen Optimierungs- und Evaluationsmessungen bestand im Störgeräuschpegel. Die Messungen zur Optimierung der Testlisten wurden mit einem Störgeräuschpegel von 60 dB SPL durchgeführt (*Wagener* et al. 1999a). Bei der Evaluation wurde jedoch mit einem für die Evaluierung von Satztests üblichen Störschallpegel von 65 dB SPL gemessen. Dies kann zusätzlich zu dem unterschiedlichen Trainingsgrad der Versuchspersonen ein Grund für die Differenz des L_{50} von Vor- und Evaluationsmessungen von ca. 1,3 dB S/N sein. Jedoch ist die Sprachverständlichkeit im Wesentlichen vom dargebotenen Signal-Rausch-Verhältnis abhängig, die Lautstärke des Störgeräuschs sollte lediglich einer »mittellauten« Lautheitsempfindung der Probanden entsprechen, dies trifft auf beide Störgeräuschpegel zu, so daß diesem Effekt keine größere Rolle zugewiesen wird. In weiteren Messungen soll dieser Effekt sowie der Einfluß beim Verwenden unterschiedlicher Störgeräusche untersucht werden.

Als Konsequenz aus dem Trainingseffekt empfiehlt sich für die praktische Anwendung des Tests die Durchführung von bis zu vier Trainingslisten. Bei einer angestrebten Meßgenauigkeit von 1 dB S/N reicht die Darbietung einer Trainings-Testliste aus, was einem Zeitaufwand von ca. 2 min entspricht.

Der Friedman-Test zeigte die Äquivalenz der Testlisten, die aufgrund der Pegelangleiche (*Wagener* et al. 1999a) und dem verwendeten Sprachmaterial (*Wagener* et al. 1999b) erwartet wurde.

Durch den Oldenburger Satztest werden daher 10 perzeptiv gleichwertige Testlisten à 10 Sätze und zwei Übungslisten bereitgestellt. Gerade für adaptive Messungen empfiehlt sich jedoch die Verwendung von 30 Sätzen pro Liste. Hierdurch läßt sich eine angestrebte Genauigkeit von 0,5 dB S/N (Standardabweichung des L_{50}) gewährleisten (*Brand* 1998; *Brand* und *Kollmeier* 1996). Dies gilt auch, wenn die Steigung der Diskriminationsfunktion aufgrund von Schwerhörigkeit der Versuchsperson auf bis zu 10 %/dB absinken sollte. Bei noch flacheren Diskriminationsfunktionen, wie sie bei hochgradig Schwerhörenden auftreten können, verschlechtert sich die Genauigkeit, so daß gegebenfalls mehr als 30 Testsätze zur Messung erforderlich sein können.

Aufgrund der Äquivalenz der 10 Testlisten können 120 verschiedene gleichwertige Tripellisten durch die Kombination der 10 Testlisten erzeugt werden, so daß genügend Testmaterial zur Verfügung steht.

Um den L_{50} und die Steigung in der Praxis durch Messen von Testlisten mit konstantem Signal-Rausch-Verhältnis zu bestimmen, muß im Idealfall eine Testliste kurz unter- und eine etwas oberhalb des L_{50} gemessen werden. Da der L_{50} bei der Messung nicht bekannt ist, benötigt diese Meßweise eine große Erfahrung des Audiometristen, damit nicht zuviele Testlisten an unterschiedlichen Signal-Rausch-Verhältnissen gemessen werden müssen. Demgegenüber haben adaptive Meßverfahren den Vorteil, daß ein Steueralgorithmus den jeweiligen Darbietungspegel des Satzes aufgrund der vorherigen Antwort der Versuchsperson auswählt (*Brand* und *Kollmeier* 1996). Das hat eine große Zeitersparnis und eine höhere Meßgenauigkeit zur Folge. Der Oldenburger Satztest hat zudem im Gegensatz zum Götttinger Satztest

den Vorteil eines hohen *j* Faktors, d. h. pro Zeiteinheit werden mehr unabhängige Test-Items getestet. Somit ist der Oldenburger Satztest besonders gut für adaptive Meßverfahren geeignet.

Schlußfolgerungen

Die in diesem Beitrag beschriebenen Evaluationsmessungen bestätigen alle in *Wagener* et al. (1999b) dargestellten Erwartungen, die aus theoretischen Berechnungen der Eigenschaften des Tests auf Basis der Optimierungsmessungen resultierten.

1. Es zeigte sich ein für die Praxis relevanter Gewöhnungs- und Trainingseffekt von maximal 2 dB S/N (Verringerung des L_{50} während der adaptiven Trainingsmessungen), der im wesentlichen über die ersten beiden 20iger Listen hinweg stattfindet. Bei einem zusätzlichen Zeitaufwand von ca. 5 min pro Messung kann und sollte der verfälschende Einfluß des Lerneffekts verringert werden.

2. Der Unterschied im mittleren L_{50} zwischen den Messungen der wortspezifischen Diskriminationsfunktionen (*Wagener* et al. 1999a) und den Evaluationsmessungen von 1,3 dB S/N (L_{50} = -8,4 dB S/N bzw. -7,1 dB S/N) kann durch das quasi-geschlossene Testverfahren der ersten Messungen (mit trainierten Versuchspersonen) nur teilweise erklärt werden. Der Unterschied ist größtenteils durch den unterschiedlichen Trainingsgrad und die verschiedene »Hörerfahrung« der Versuchspersonen zu erklären. Eventuell haben zusätzlich die unterschiedlichen Störgeräuschpegel einen Einfluß auf die Meßergebnisse. Die Abhängigkeit des L_{50} vom Störschallpegel sollte daher in Zukunft für dieses Sprachmaterial untersucht werden, jedoch ist laut Literatur keine große Abhängigkeit zu erwarten.

3. Die experimentell gefundene Steigung der Testlisten von 17,1 %/dB stimmt mit der für diesen Test theoretisch möglichen (17,2 %/dB, aus *Wagener* et al. 1999a) überein. Diese hohe Steigung ermöglicht eine effiziente Bestimmung der Sprachverständlichkeit im Störgeräusch. Die geringen Standardabweichungen von L_{50} und *m* der einzelnen Testlisten sowie der Friedman-Test zeigt die hervorragende Homogenität des Testmaterials.

4. Die 10 zusammengestellten Testlisten können zu 120 unterschiedlichen Tripellisten zusammengestellt werden, die Forderung nach einer hohen Anzahl an Testlisten ist somit erfüllt. Zusätzlich können die Listen auch wiederholt gemessen werden, da sie aufgrund der semantisch nicht vorhersagbaren Struktur nicht im Gedächtnis behalten werden können. Schon während der Optimierungsmessungen wurde die direkt aufeinanderfolgende Darbietung derselben Testliste selbst bei überschwelligen Pegeln von den Versuchspersonen nicht bemerkt.

5. Die Vorhersagbarkeit der Sätze ist sehr gering (j = 4,3 bei -5 dB S/N), so daß bei der Messung eines Trials ein hoher Informationsgewinn stattfindet. Daher eignet sich der Oldenburger Satztest besonders für adaptive Meßverfahren.

6. Die in *Wagener* et al. (1999b) formulierten Anforderungen an den Satztest bezüglich der Durchführbarkeit im Störgeräusch, der Steilheit der Diskriminationsfunktion sowie der Anzahl und Wiederholbarkeit der Testlisten werden nach den Ergebnissen dieses Beitrags vom Oldenburger Satztest erfüllt.

Unterstützt von der DFG, KO 942/13-1.
Vielen Dank an A. Gorges und an die Probanden für die Durchführung der Messungen.

Der Oldenburger Satztest ist erhältlich über das Hörzentrum Oldenburg, c/o Universität Oldenburg, Carl von Ossietzky-Str. 9-11, 26111 Oldenburg, Tel: 0441 973 8997, Fax: 0441 973 8998.

Corrigendum

In der »Zeitschrift für Audiologie« 2/99, auf Seite 47 des Originalbeitrages »Entwicklung und Evaluation eines Satztests für die deutsche Sprache – Teil II: Optimierung des Oldenburger Satztests« von Kjrsten Wagener, Thomas Brand und Birger Kollmeier (Universität Oldenburg) wurde aufgrund einer technischen Panne eine Gleichung an der falschen Stelle plaziert.
Gleichung (7) muß wie folgt lauten:

$$m_{ges} = \frac{m_{Wort}}{\sqrt{1 + \frac{\sigma_j^2}{\sigma_{Wort}^2}}}$$

Literaturverzeichnis

Boothroyd A, Nittrouer S (1988) Mathematical treatment of context effects in phoneme and word recognition.
J Acoust Soc Am 84 (1), 101-114

Brand T (1994) Effiziente Bestimmung psychometrischer Funktionen mit Sprachverständlichkeitstests.
Diplomarbeit. Georg-August-Universität Göttingen

Brand T (1998) Persönliche Mitteilung

Brand T, Kollmeier B (1996) Adaptive Testverfahren in der Audiologie. Fortschritte der Akustik - DAGA 96, 60–63

Hagerman B (1984) Some Aspects of Methodology in Speech Audiometry.
Dissertation. Karolinska Institutet Stockholm, Schweden

Hagerman B (1996) Persönliche Mitteilung

Hagerman B, Kinnefors C (1993) Efficient adaptive methods for measurements of speech reception thresholds in quiet and in noise.
Karolinska Institutet, Teknisk Audiologi, Stockholm

Kollmeier B, Müller C, Wesselkamp M, Kliem K (1992) Weiterentwicklung des Reimtests nach Sotscheck.
In: Kollmeier B (Hrsg.) Moderne Verfahren der Sprachaudiometrie. Median-Verlag, Heidelberg. 216–237

Kollmeier B, Wesselkamp M (1997) Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment.
J Acoust Soc Am 102 (4), 2412-2421

Niemeyer W (1967) Sprachaudiometrie mit Sätzen I: Grundlagen und Testmaterial einer Diagnostik des Gesamtsprachverständnisses.
HNO 15, 335-343

Sachs L (1992) Angewandte Statistik.
Springer-Verlag, Heidelberg

Wagener K, Brand T, Kollmeier B (1999a) Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests.
Z Audiol 38 (2), 44-56

Wagener K, Kühnel V, Kollmeier B (1999b) Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests.
Z Audiol 38 (1), 4-15